KSU Team's Dialogue System at the NTCIR-13 Short Text Conversation Task 2

Yoichi Ishibashi, Sho Sugimoto, Hisashi Miyamori (Kyoto Sangyo University, Japan)
{g1445539, g1444693, miya}@cc.kyoto-su.ac.jp

Abstract

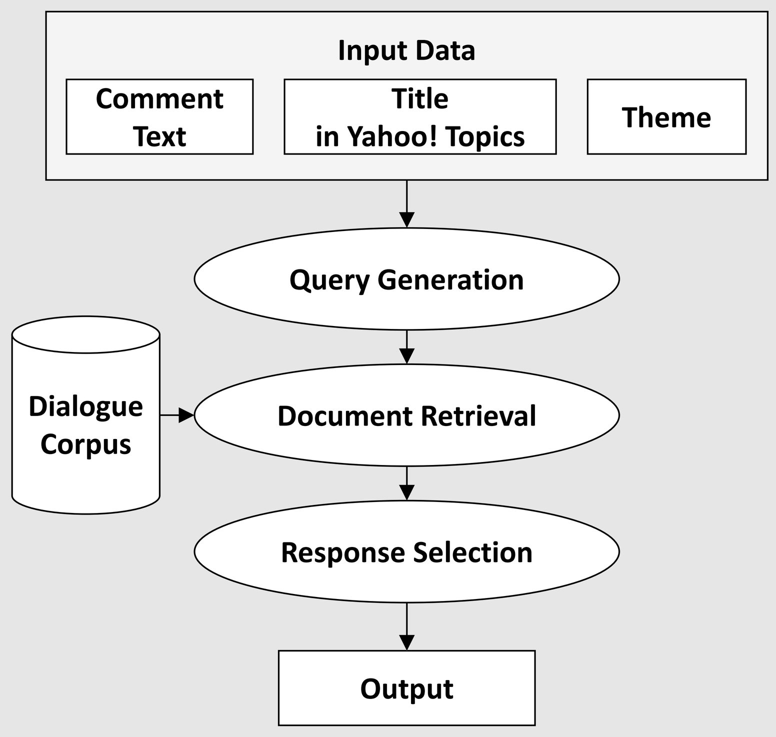
In the retrieval-based methods, a comment text with high similarity with the given utterance text is obtained, and the reply text to the comment text is returned as the response to the input.
In the generation-based method, we propose the Associative Conversation Model that generates visual information from textual information and uses it for generating sentences in order to utilize visual information in a dialogue system without image input.

Conclusion

In the retrieval-based methods, it was found that the accuracy is improved by excluding place names and person names from article themes.
In the generation-based method, we could not find enough evidence from the results using the evaluation data of STC to show that the associated visual information would accelerate to generate more appropriate responses. Newly, we conducted additional experiment based on the problems found in the evaluation results. As a preliminary result, it was confirmed that **visual information seemed to work effectively** in several examples.

Retrieval-based method

Overview of Architecture



Query type	Weight
Comment text	1.0
Title in Yahoo! Topics	9.1
Theme	3.5

- **Query Generation**
  A query used for similarity search is generated from the given input data by using the analyzers provided in Solr.
  In Run 1, the place names and person names were removed from the words in the theme as the search query.
- **Document Retrieval**
  Similarity search by Okapi BM25 is conducted for only comment texts in the Yahoo! news comments data.
  The each query was weighted.
  (refer to the table on the left)
- **Response Selection**
  The final output is the reply text corresponding to the obtained comment text in Document Retrieval.

Result

Input	<Comment text> I stopped going out by bicycle because the forecast said it would rain or snow at 12 o'clock, but there is nothing falling yet.	
	<Title in Yahoo! Topics> Snow falls in downtown Tokyo. Snow likely to pile up in the Kanto flat land	
	<Theme> **Tochigi Prefecture**, Weather forecast, **Gunma prefecture**, **Ibaraki prefecture**, Snow damage and its measures	
	Run 1	Run 3
Rank 1	It would be fun if the snow occasionally piled up in downtown area. Because even though it is the Kanto local news, the news program of the key station makes a fuss like a nationwide incident.	It's amazing that as many as two schools will participate from Kochi prefecture, including a "21st-century selection" slot.
Rank 2	Unusually, there was no snow last month, wasn't it?	It's also heavy snow in the southern part of Kanagawa prefecture.
Rank 3	It'd be easier to objectively understand the situation with the expressions like possibility or probability. Ridiculous to describe it with the word "worry" ... It's quite an emotional expression.	In the electricity industry, Niigata prefecture is within the jurisdiction of Tohoku Electric Power Co.

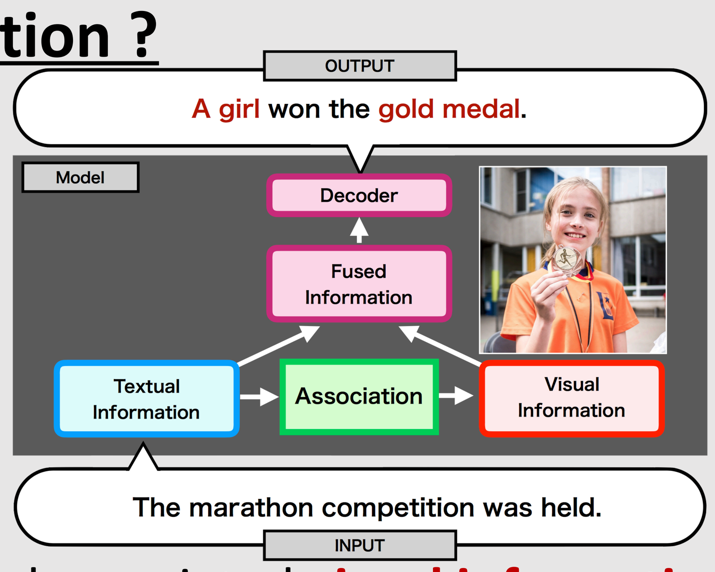
Generation-based method

Overview

Why do we use the visual information ?

→ The image or video may contain more detailed information than texts.

How can the visual information be used without image input ?



1. **Visual association** from the **input text**
2. **Fusion** between **textual information** and associated **visual information**
3. Response text generation based on the **fused information**.

Learning Method

Prior train 1: Extraction of context vectors between textual and visual information
The model has two LSTM encoder, one fusion layer, and one decoder with attention. C_t is calculated by following equation at the fusion layer. The trained model is used to extract the correspondence between the textual and visual information.

$$C_t = W_{txt}C_t^{txt} + W_{vis}C_t^{vis} + b_c$$

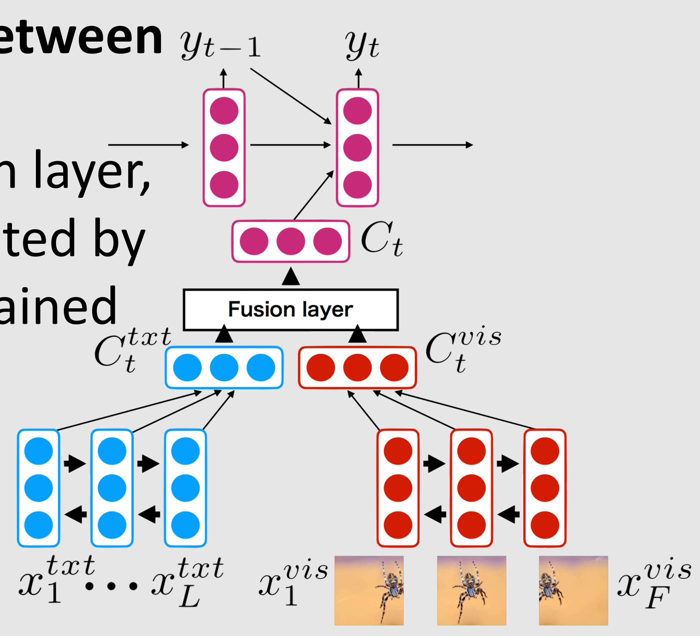


Figure 1: Network in step 1

Prior train 2: Learning for visual association
The associative encoder (LSTM) is trained, which inputs the textual context vector C^{txt} extracted in step 1 and outputs the visual context vector C^{vis} corresponding to the input utterance text.

$$C_t^{vis} = LSTM(C_t^{txt})$$

Final step: Generation of response text via association
Figure 2 shows a network used in step 3. Learning is performed in the network where the visual encoder in step 1 is replaced with the associative encoder trained in step 2.

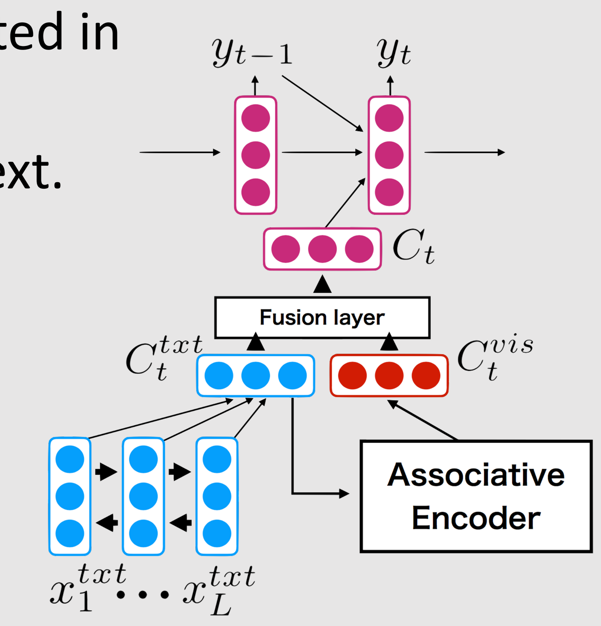


Figure 2: Network in step 3

NTCIR-13 STC-2 formal-run

Our generation-based model (Run2) is trained with **TV drama data**.

TV drama Dataset

- Subtitles in TV drama and the video where the subtitles were displayed.

Results

- In Rule 1, Mean Acc_{L2}@2 and Mean Acc_{L2}@1, which are the accuracy when only L2 is correct, were more than 0.5, indicating that Run2 had many L1 and few L2.
- In Rule 2, Mean Acc_{L2}@2, and Mean Acc_{L2}@1 got worse than those in Rule1
  The factors can be considered as problems;
  ① insufficient amount of learning data, ② lack of tuning, and ③ many unknown words (proper nouns) in the test data.

Additional Experiments

- Comparison between our model and baseline model
  Baseline model: An encoder-decoder model with attention mechanism
- Our generation-based model is trained with **TV news data** instead of TV drama data.

TV news Dataset

- Subtitles in TV news and the video where the subtitles were displayed.

Results and Analysis of Visual Association

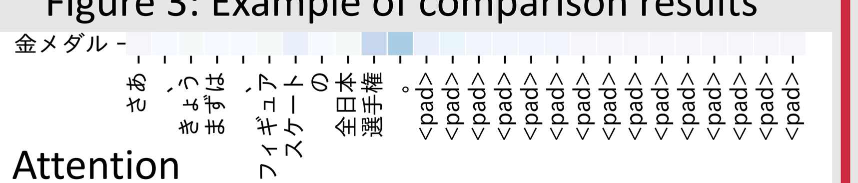
- Our model (ACM) associated visual information from the input text.
- Our model could generate responses including useful information compared with the model without association.

Example of Visual Association

- our model associates the snow scene with the input text and generated the word **"snowy"** (Fig.3 left).
- Note that it is a fact that it actually snowed on the day of the exam.

Input	The University Entrance exam will be held on 14th and 15th.	Well, today is All Japan Figure Skating Championships.
	14日と15日は、大学入試センター試験です。	さあ、きょうまずは、フィギュアスケートの全日本選手権。
Output by Baseline	There will be a large-scale fire that is also in western Japan and eastern Japan.	Aiming for four consecutive championships in the women's singles, athletes of the Japanese championships participated in the tournament.
	西日本や東日本にもなる、も大規模な火災ができます。	女子シングルで4連覇を狙うで、日本選手権の選手が、大会に出場しました
Output by ACM	It is highly expected to be **snowy** and windy.	A player who has won the **gold medal** in women's singles.
	雪の風の予想が広がっています。	女子シングルで3した**金メダル**を獲得している選手。
Image associated from the input (Source is described below)		
Words generated mainly from the associated image	**Snowy**	**Gold medal**
	雪	金メダル

Figure 3: Example of comparison results



Attention

Source: *Fig3 left: "NHKニュース7", NHK, 11 January 2017* *Fig3 right: "NHKニュース7", NHK, 23 February 2017*

京都産業大学